

## Statistics Final Exam

Consider the following analysis of a dataset from economics. Each line (observation) in the dataset corresponds to a fortune 500 company. The analysis concerns the following variables:

- *lsales* – the natural log of the company’s sales
- *lassets* – the natural log of the company’s assets
- *lmarket.value* – the natural log of the company’s current market value
- *lemployees* – the natural log of the number of employees of the company
- *energy* – an indicator variable, 1 if the company is in the energy sector, 0 otherwise

Here is some of the *R* session for the analysis:

```
> m1 <- lm(lsales ~ lassets + lmarket.value + lemployees + energy)
> summary(m1)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	3.94648	0.45925	8.593	8.85e-13
lassets	0.19156	0.05282	3.627	0.00052
lmarket.value	0.07861	0.06783	1.159	0.06015
lemployees	0.60392	0.05462	11.056	0.03824
energy	0.61879	0.14339	4.316	4.86e-05

Residual standard error: 0.4833 on 74 degrees of freedom

Multiple R-squared: 0.845, Adjusted R-squared: 0.8388

F-statistic: 136.3 on 3 and 75 DF, p-value: < 2.2e-16

1. Based on this model, what would be the effect of a one point increase in *lassets* on *lsales*?

A one point increase in *lassets* will result in an increase of .19156 in *lsales*.

2. Based on this model, what would be the effect of a one point increase in *lassets* on sales?

Since *lsales* is the logarithm (base  $e \approx 2$ ) of sales, a roughly five point increase in *lassets* will increase *lsales* by one point, and approximately double *sales*.

3. Estimate a 95% confidence interval for  $\beta_{lassets}$ .

$$0.19156 \pm 2 \cdot 0.05282 = (0.13874, 0.24438)$$

4. What is the sample size, and how do you know?

Residual standard error on 74 degrees of freedom, plus one each for (*Intercept*), *lassets*, *lmarket.value*, *lemployees*, and *energy* equals 79.

5. What are the null and alternative hypotheses being tested by the p-value .03824?

$$\begin{aligned} H_0 : \beta_{lemployees} &= 0 \\ H_1 : \beta_{lemployees} &\neq 0 \end{aligned}$$

6. What is the conclusion of the test above?

Reject  $H_0$  with a  $p$ -value of 0.03824 (less than 0.05) and conclude that  $\beta_{lemployees} \neq 0$ .

7. What are the two models being tested above?

$$\begin{aligned} H_0 : lsales &= \beta_0 + \beta_1 lassets + \beta_2 lmarket.value + \beta_4 energy + \epsilon \\ H_1 : lsales &= \beta_0 + \beta_1 lassets + \beta_2 lmarket.value + \beta_3 lemployees + \beta_4 energy + \epsilon \end{aligned}$$

8. What are the null and alternative hypotheses being tested by the  $p$ -value  $< 2.2\text{e-}16$ ?

$$H_0 : \beta_{\text{lassets}} = \beta_{\text{lmarket.value}} = \beta_{\text{lemployees}} = \beta_{\text{lenergy}} = 0$$

$H_1$  : at least one of the  $\beta$ s above is non-zero

9. What is the conclusion of the test above?

With a  $p$ -value of nearly zero, we reject  $H_0$  and conclude that at least one of the slopes is non-zero (at least one of the variables in our model has a linear effect on *lsales*).

10. What are the two models being tested above?

$$H_0 : \text{lsales} = \beta_0 + \epsilon$$

$$H_1 : \text{lsales} = \beta_0 + \beta_1 \text{lassets} + \beta_2 \text{lmarket.value} + \beta_3 \text{lemployees} + \beta_4 \text{lenergy} + \epsilon$$

11. How many of the confidence intervals computed by the command:

```
> confint(m1)
```

would contain 0 (which ones, if any, and why)?

Only the one for the *lmarket.value* variable because the  $p$ -value is larger than .05.

12. What percentage of the variability in *lsales* is explained by the model?

84.5% (from the multiple  $r^2$ ).

13. What is the model for energy sector employees? (Write out the actual numerical coefficients for the  $\beta_i$ . Feel free to round to two decimal places if you like.)

For energy sector employees *energy* is equal to 1, so we have:

$$\begin{aligned} \text{lsales} &= 3.95 + .19 \text{lassets} + .08 \text{lmarket.value} + .60 \text{lemployees} + .62 \text{lenergy} + \epsilon \\ &= 3.95 + .19 \text{lassets} + .08 \text{lmarket.value} + .60 \text{lemployees} + .62 \cdot 1 + \epsilon \\ &= (3.95 + .62) + .19 \text{lassets} + .08 \text{lmarket.value} + .60 \text{lemployees} + \epsilon \\ &= 4.57 + .19 \text{lassets} + .08 \text{lmarket.value} + .60 \text{lemployees} + \epsilon \end{aligned}$$

14. Suppose that after the exam, I pass a collection plate and our class uses the proceeds to purchase a U.S. automobile manufacturer. Our ability to secure governmental bailout funding will be proportional to our potential sales, and hence a need to predict future sales.

```
> predict(m1,newdata=our.co,interval='confidence',level=.95)
```

```
      fit      lwr      upr
8.725756 8.5666 8.88491
```

- (a) What does it mean that we are 95% confident with this 95% confidence interval?

If we were to repeat this sampling many times (say 100) and compute exactly the same model each time with the new data, then about 95% of these models would generate a confidence interval that contains the exact answer. We hope that our one sample is one of these correct ones.

- (b) What, exactly, is it that we are 95% confident that this 95% confidence interval contains?

We are 95% confident that this confidence interval contains the average *lsales* for the population of all companies with *lasset*, *lmarket.value*, etc. exactly the same as our company. This is almost certainly not the correct type of confidence interval for this situation as we are interested in the sales (*lsales*) of our one company that we are going to purchase, not the average *lsales* of the population of all similar companies.

15. For each of the diagnostic plots listed below, tell what, exactly, is being (informally) tested for and the implications of the specific problems that might be found by that plot for future analysis.

- (a) The residual plot.

We want to see a nice random scatter about the  $x$ -axis. We are looking for heteroskedasticity (this would affect the reliability of our confidence intervals like the ones produced above), or any sort of pattern in the plot, say a curve or an upward trend (implying a mis-specified model), or perhaps we could see some outliers (see below). We don't see any great indicators of heteroskedasticity, an only perhaps some mild outliers (between  $x = 6$  and  $x = 7$  up at the top).

- (b) The jackknife residual plot.

We are looking for outliers. Hopefully all of the points are not too far away from the  $x$ -axis. Outliers can indicate that there might be some lurking variable (like the red dwarf stars in the *star.data* data set) or they might be influential points affecting the estimates in our model, or that we might be wrong in our predictions and confidence intervals. We have no outliers here, at least not according to the Bonferroni correction (which is actually quite conservative), but 72 and 75 are close, and might need investigating anyway.

(c) The hat value plot.

High leverage points. Leverage in and of itself is not bad, but high leverage points do have the potential to pull a regression line toward themselves, affecting the coefficients in our model. The points 25, 33, 40, 44, and 73 are identified by this test.

(d) The diffits plot.

Here we are looking for influential points. If deleting a point from the model makes a dramatic difference in the predictions made, we need to investigate what it is about that point that makes such a difference, and whether predictions are more or less reliable with its inclusion in the model. Here we see lots of points above the cutoff line, so if the purpose of this modeling is prediction, we might be in trouble.

(e) Collectively, the dfbetas plots.

Here we are again looking for influential points, but the influence is on the terms (coefficients, slopes, whatever you want to call them) rather than the predictions made by the model. Particularly for the *lassets* variable there are many influential points. Companies 19 and 73 show up in three of the four deletion diagnostic plots here.

16. Looking at the added variable plot for *lemployees*, make predictions for the output of the summary command below. In particular comment on what you think will likely change (relative to *m1*) regarding  $r^2$ , the statistical significance of the slopes, and on the  $p$ -value for the slope of *lemployees*<sup>2</sup>.

```
> m2 <- lm(lsales ~ lassets + lmarket.value + lemloyees +  
I(lemloyees2) + energy)  
> summary(m2)
```

While squinting just right at the added variable plot might possibly indicate some sort of mild heteroskedasticity, there is nothing in the plot indicating that there is any sort of curvilinear relationship going on here. Thus while  $r^2$  is sure to go up with the addition of another variable, the adjusted  $r^2$  will probably not, and the  $p$ -value for *lemployees*<sup>2</sup> is likely to be high.

17. The analyst decides to explore a further model (*m3*). What are the null and alternative hypotheses being tested by the *anova* command?

```
> m3 <- lm(lsales ~ lassets + lemloyees + lmarket.value + energy  
lassets*energy + lmarket.value*energy + lemloyees*energy)  
> anova(m1,m3)
```

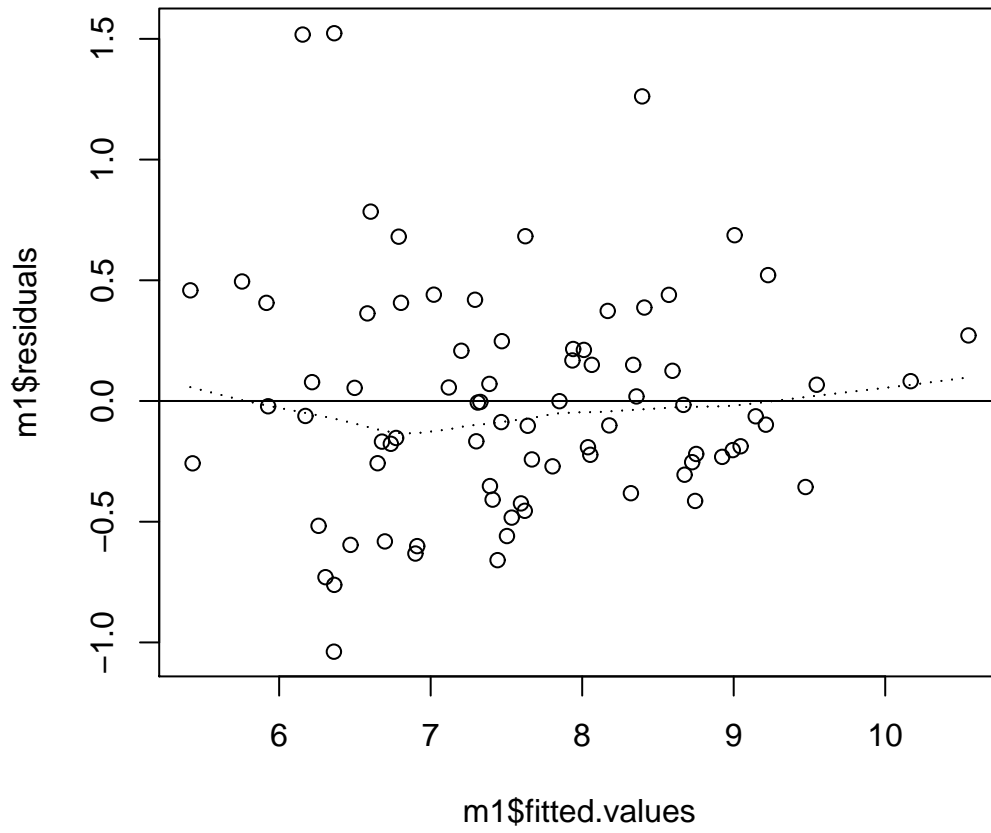
	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	75	13.9979				
2	71	11.9582	4	2.0397	3.0276	0.02304

$H_0$  :  $m1$  and  $m3$  are equivalent in their explanatory value  
 $\beta_{lassets*energy} = \beta_{lmarket.value*energy} = \beta_{lemployees*energy} = 0$   
 $H_1$  :  $m3$  is better than  $m1$   
at least one of the  $\beta$ s above is non-zero.

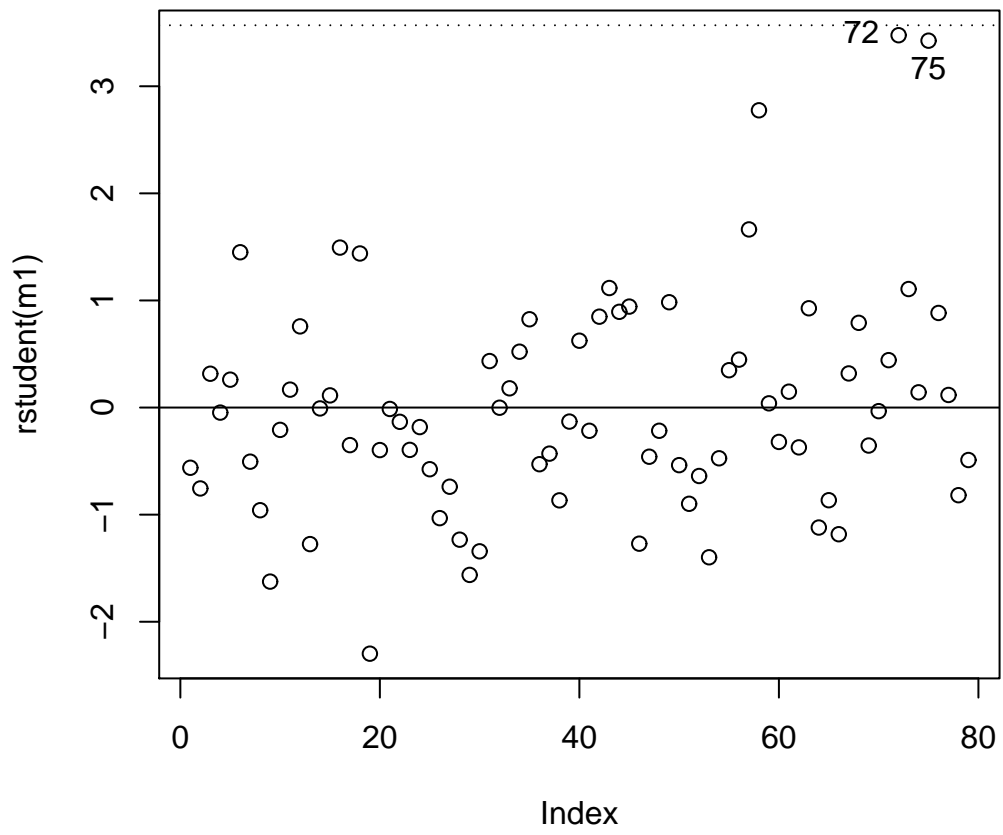
18. What is the result of this test? What does that mean about the relationship between a firm's being in the energy sector, its *lassets*, *lemployees*, *lmarket.value*, and the  $y$ -variable *lsales*?

With a  $p$ -value of .02 (less than .05) we reject  $H_0$  in favor of the alternative, and conclude that there is a different effect of *lassets*, *lmarket.value*, and *lemployees* on *lsales* when looking at energy sector firms when compared to other types of companies.

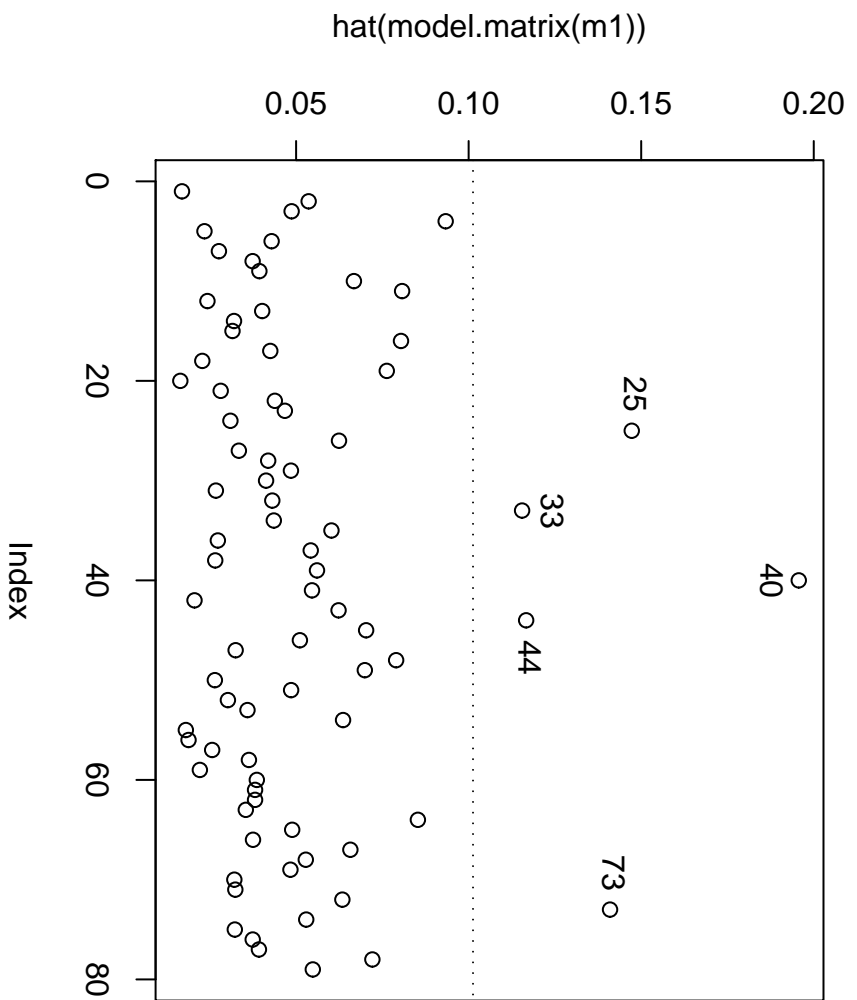
m1 residual plot



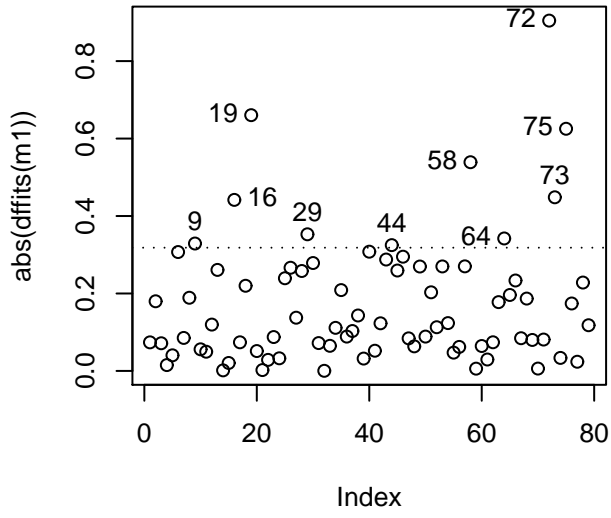
m1 jackknife residual plot



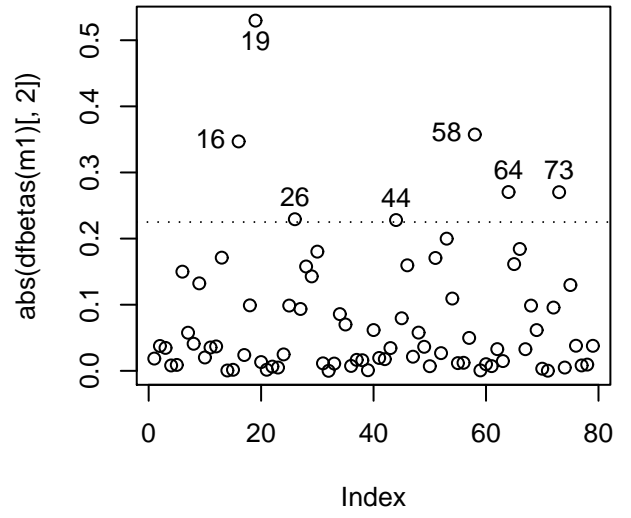
**m1 hat plot**



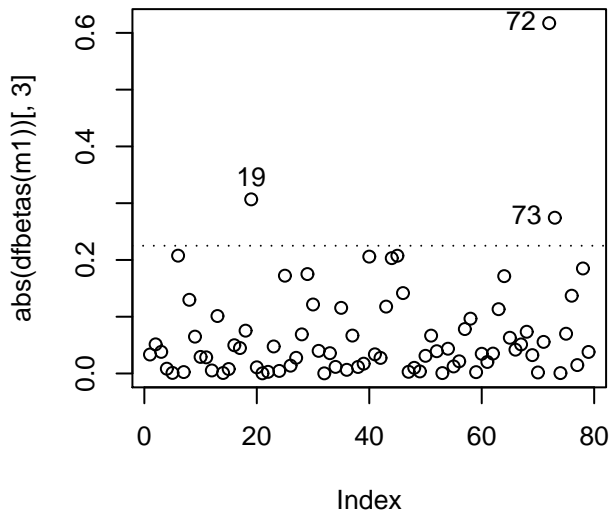
**dffits m1**



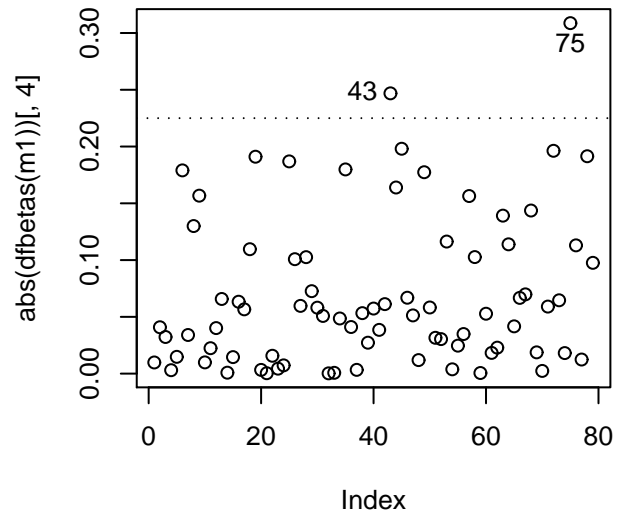
**lassets dfbetas**



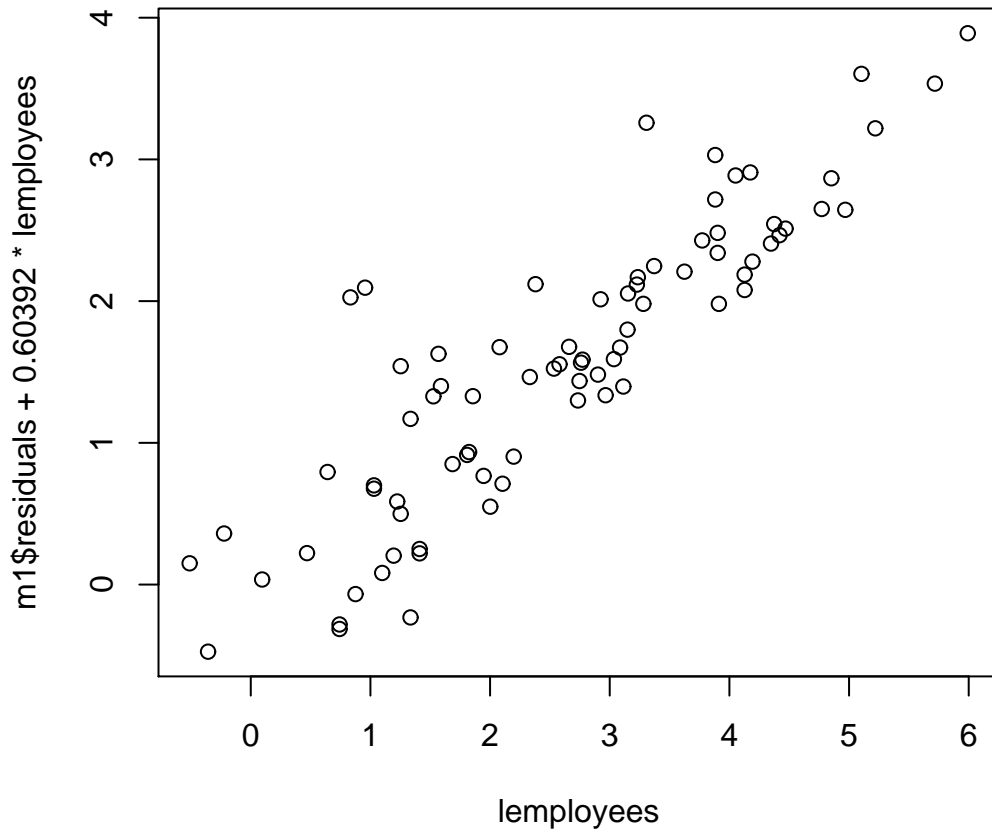
**lmarket.value dfbetas**



**lemployees dfbetas**



### employees added variable plot



```
> m1 <- lm(lsales ~ lassets + lmarket.value + lemloyees + energy)
> summary(m1)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	3.94648	0.45925	8.593	8.85e-13
lassets	0.19156	0.05282	3.627	0.00052
lmarket.value	0.07861	0.06783	1.159	0.06015
lemloyees	0.60392	0.05462	11.056	0.03824
energy	0.61879	0.14339	4.316	4.86e-05

Residual standard error: 0.4833 on 74 degrees of freedom

Multiple R-squared: 0.845, Adjusted R-squared: 0.8388

F-statistic: 136.3 on 3 and 75 DF, p-value: < 2.2e-16

```
> predict(m1,newdata=our.co,interval='confidence',level=.95)
```

	fit	lwr	upr
	8.725756	8.5666	8.88491

```
> m3 <- lm(lsales ~ lassets + lemloyees + lmarket.value + energy
           lassets*energy + lmarket.value*energy + lemloyees*energy)
```

```
> anova(m1,m3)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	75	13.9979				
2	71	11.9582	4	2.0397	3.0276	0.02304